FF0082

Prefinishing at the Genome Sequencing Center

Amy Reily, Robert S. Fulton and Richard K. Wilson

Genome Sequencing Center, Washington University School of Medicine St. Louis, MO 63108

Abstract: The Prefinishing group at the Genome Sequencing Center (GSC) offers a convenient, efficient, and cost-effective method for improving the quality of shotgun sequence data. The GSC uses Autofinish (Gordon, D., University of Washington) to improve low quality regions, close gaps, or both in a variety of assemblies. The Prefinishing process is used on BAC clones as well as whole genome assemblies of varying sizes (small bacterial assemblies, or large assemblies such as the chimpanzee and chicken genomes), or regions of either BACs or whole genome assemblies such as gene regions or non-repeat regions. The GSC's prefinishing pipeline is highly successful in closing gaps and improving low quality regions in all projects. For the clone-based Maize project, the number of contigs is reduced by 50% during the prefinishing process. For NHGRI BACs, the results are even better. This prefinishing process can stand alone as an improvement to BACs or whole genome assemblies, or can significantly reduce finishing time and costs for projects scheduled for additional improvement efforts.

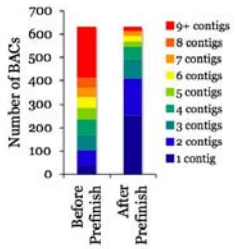
GSC Prefinishing Process

- Assembly is examined for misassemblies
- AutoEdit is run to tag editable, low consensus quality areas
- Autofinish is run
  - Calls reactions for low consensus quality and gaps
  - Uses only templates available in the freezers
  - All reactions are oligo walks using 1 or 2 templates per oligo
- Reactions processed through automated pipeline
- Project is reassembled (using phrap or pcap)
- Projects can then get a second round of Autofinish, get passed to a finishing group, or be complete at this stage

Prefinish Results for NHGRI BACs

NHGRI BACs are shotgunned to 8-10x coverage and then receive 2 rounds of Autofinish before being passed on to finishing. These BACs are submitted as contiguous projects with no low quality regions.

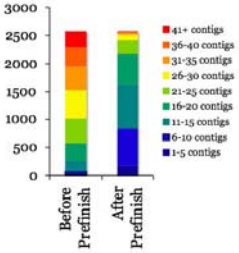
- Two rounds of Autofinish are done on each BAC
- PCR is ordered for any unspanned gaps
- In addition to reduction in contig number, many low quality regions are resolved or improved
- Average contigs before prefinish=6, average contigs after prefinish=2.5
- Prefinishing is done on BACs from many different organisms
  - Chimpanzee
  - Chicken
  - Mouse
  - Orangutan
  - Macaque
  - Platypus



Prefinish Results for Maize BACs

Maize BACs are shotgunned to 4-6x coverage and then receive 2 rounds of Autofinish before being passed to finishing. Only the gene regions of these clones are finished. For more information see poster FF0081.

- Two rounds of Autofinish are done on each maize BAC
- Regions of repeat are tagged with doNotFinish tags to prevent Autofinish from calling reactions on those areas
- Reactions are called for all spanned gaps regardless of repeat tagging
- PCR is called for missing BAC ends
- Average contigs before prefinish=28, average contigs after prefinish=15

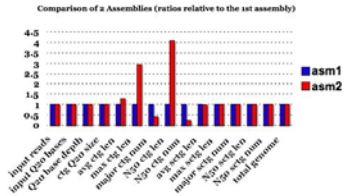


Prefinishing Capabilities

- BAC clones
  - Prefinishing on entire clone for improvement before finishing
  - Prefinishing only for improvement of clones not being finished
  - Autofinish on specified regions of clone (gene-regions, non-repeat regions, specific contigs)
  - Can exclude contigs within certain sizes, contamination contigs, or repeat-only contigs from Autofinish
- Small whole genome assemblies (genomes less than 10Mb)
  - Autofinish as a precursor to finishing or manual sequencing improvement
  - Prefinishing alone can quickly and efficiently improve these genomes
- Large whole genome assemblies (genomes greater than 10Mb)
  - Autofinish can be run on entire genomes
  - Autofinish can be limited to specific regions (exons, gene-regions, non-repeat regions, regions not finished in BACs)
  - Can exclude contigs from Autofinish based on size, contamination, or repeats
  - Can select a small region of the genome, break it into its own ace file, and run Autofinish on just that area

Prefinishing Bacterial Projects

- Prior to Autofinish the databases are sorted by a finisher who resolves misassembly and makes obvious joins.
- Autofinish is then run using 2 templates per oligo for each gap and low quality area. AutoEdit is not used to tag editable regions in these projects.
- After one round of Autofinish the project is reassembled and enters the manual sequence improvement queue.
- Shown below is a graph comparing the status of two assemblies of actinomyces odontolyticus asm1 is the draft assembly, asm2 is the assembly after one round of Autofinish reactions.



Prefinishing the Chimpanzee Genome

- Pcap whole genome assembly broken into 400 ace files for Autofinish
- Ace files tagged with doNotFinish tags on regions covered by finished BACs and repeat regions
- AutoEdit is run to tag editable, low consensus quality areas
- Autofinish run only on contigs larger than 20 reads and 5Kb
  - Called reactions for low consensus quality and gaps
  - Called reactions only on fosmid templates
  - Called only oligo walks using 1 template per oligo
- A total of 378, 093 reactions were called by autofinish
- Thus far, 32,733 reactions are completed